Is regularization necessary? A Wald-type test under non-regular conditions

Maastricht University

Artem Duplinskiy

Is regularization necessary? A Wald-type test under non-regular conditions

RM/14/025

GSBE

Maastricht University School of Business and Economics
Graduate School of Business and Economics

P.O Box 616
NL- 6200 MD Maastricht
The Netherlands

Is regularization necessary? A Wald-type test under non-regular conditions.*

Artem Duplinskiy†

Maastricht University, SBE, Department of Quantitative Economics

July 2, 2014

Abstract

We study hypotheses testing in the presence of a possibly singular covariance matrix. We propose an alternative way to handle possible non-regularity in a covariance matrix of a Wald test, using the identity matrix as the weighting matrix when calculating the quadratic form. The resulting test statistic is not pivotal, but its asymptotic distribution can be approximated using bootstrap methods. In order to prove the validity of the approximations, we show that the square root of a positive semi-definite matrix is a continuously differentiable transformation with respect to the elements of the matrix. This result is important for the continuous mapping theorem to be applicable. We use two types of approximations. The first uses the parametric bootstrap and draws from the asymptotic distribution of the restriction with an estimated covariance matrix. The second applies the residual bootstrap to obtain the distribution of the test and delivers critical values, which control size and show good empirical power even in small samples. In contrast to regularization approaches, the test statistic considered in this paper does not involve arbitrary truncation parameters for which no practical guidelines are available and does not modify the information in the data.

JEL Codes: C12; C15; C32.
Keywords: Nonlinear restrictions; Wald test; Multiple time series; Regularization.

*We thank Francisco Blasques, Thomas Götz, Stephan Smeekes, Hanno Reuvers, Franz Palm, Jean-Pierre Urbain and the participants of the 7th International Conference on Computational and Financial Econometrics in London for useful suggestions and comments on earlier versions of the paper.

†Correspondence to: Artem Duplinskiy, Maastricht University, School of Business and Economics, Department of Quantitative Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: a.duplinskiy@maastrichtuniversity.nl, Tel.: +31 43 388 3944, Fax: +31 43 388 20 00

1 Introduction

This paper contributes to the old problem of hypotheses testing under non-regular conditions. Already Moore (1977) uses generalized inverse (g-inverse) to construct chi-square test of goodness of fit. G-inverses are used in the goodness-of-fit tests (see e.g. Andrews, 1988a, Andrews, 1988b), as well as, to construct testing procedures for the rank deficient linear model (see, e.g. Mitra, 1980), and generalized method of moments specification tests (see, e.g Newey, 1987), among others. For more examples we refer to the introduction of Andrews (1987).

Possibility of singular covariance matrix in the computation of a test statistic complicates hypothesis testing. Two types of singularity, as discussed in Dufour et al. (2011), confront researchers. One type is called *reducible* singularity and covers the cases, where a degenerate matrix is the result of different rates of convergence for different components of the estimator. Such singularity could be removed by an appropriate rotation and rescaling of the elements. Therefore, it does not affect the asymptotic distribution of the Wald statistic (see, e.g. Hamilton, 1994, Chapter 16, pages 457-460). The second type of singularity is called *irreducible*. Different reasons may cause this type of singularity: large number of parameters relative to the number of observations, strongly correlated instruments, which lead to collinearity problems, or simply by redundant variables. *Irreducible* singularity is harmful for computation of a test statistic.[1] This paper suggests an approach to deal with this issue.

Andrews (1987) claims that the use of the generalized inverse (g-inverse) of the sample matrix instead of the g-inverse of the population matrix does not affect the asymptotic distribution of the quadratic form if the rank of the consistent estimator of the sample covariance matrix is the same as the one of the population matrix with probability one. But the asymptotic distribution of the quadratic form is modified otherwise. The paper provides necessary and sufficient conditions for the g-inverse of the sample matrix to converge to its population counterpart.

Existing methods for hypothesis testing under non-regular conditions often rely on the regularization approach, which modifies the test statistic to circumvent the problem: it either makes an estimate of a covariance matrix of a restrictions invertible or it transforms a weighting matrix. For instance, Lütkepohl and Burda (1997) suggest two procedures. One is a scheme that sets the small eigenvalues of the matrix to zero[2] to produce a consistent estimator for the rank of the population matrix and is based on the Takagi's factorization of a square, symmetric matrix (see, e.g. Horn and Johnson, 1985, Chapter 4, Corollary 4.4.4). The second approach ensures the full rank of the covariance matrix by adding a noise term. The authors obtain a full rank matrix at the cost of adding irrelevant information to the data in this case. Moreover, as pointed out by Andrews (1987), g-inverses are not necessarily continuous functions of the initial covariance matrices so the application of the continuous mapping theorem might not be justified. Dufour et al. (2011) argue that eigenvectors are not continuous functions in the elements of the matrix. Hence, the eigenvectors corresponding to eigenvalues with multiplicity larger than one are not

[1] For details on two types of singularity and various examples of singularity in the covariance matrix, see the introduction of Dufour et al. (2011).

[2] This procedure can be viewed as a pre-test, hence falling to the critique of Leeb and Pötscher (2003) and Leeb and Pötscher (2005).

uniquely defined when performing the singular value decomposition of a matrix. In light of this fact, regularization methods that perform such a decomposition may lead to incorrect distributional results because the convergence of the estimates towards their population counterparts is not guaranteed.

Dufour et al. (2011) use a new class of regularized inverses to resolve the singularity issue. They exploit total eigenprojection techniques (see, e.g, Katō, 1995), combined with the variance regularizing function (VRF) that modifies the small eigenvalues falling below a certain threshold c, so that their inverse is well defined. Under specific regularity conditions, the new regularized inverse converges to its regularized counterpart. The idea behind regularized inverses is quite similar to the first approach of Lütkepohl and Burda (1997), though there are some substantial differences. Regularization is performed via two channels: by modifying the small eigenvalues of the original matrix, and by introducing the eigenprojection techniques, which provide a way to handle eigenvalues with multiplicity higher than one. Thus, the regularized inverse is shown to be a continuous transformation for a given covariance matrix and asymptotic results follow from the continuous mapping theorem.

The regularization approaches discussed in the previous paragraphs modify the information contained in the data to satisfy the rank condition of Andrews (1987). Also the regularization techniques rely on the choice of tuning parameters, the optimal values of which is unknown to the researchers. We propose an alternative way to handle potential non-regularity in a covariance matrix of a Wald test that does not modify the information in the data, neither it depends on the truncation parameters. We consider a Wald-type test that has the identity matrix instead of the covariance matrix when calculating the quadratic form.

To a large extend, the testing approach falls into the class of Monte Carlo tests based on a consistent point estimates of nuisance parameters (Dufour, 2006). Although the test statistic is not pivotal, we approximate the asymptotic distribution of this test, using bootstrap. The idea to bootstrap the function of the population covariance matrix is not a novelty in this paper. Beran and Srivastava (1985), for example, consider bootstrap tests and confidence regions for functions of the population covariance matrix. They show that the tests under consideration have the desired asymptotic levels, provided model restrictions, such as multiple eigenvalues in the covariance matrix, are taken into account when designing the bootstrap algorithm. In an independent study, Duchesne and Francq (2014) consider the properties of the Wald test based on the identity weighting matrix and compare it to tests based on generalized and $\{2\}$-inverses.

We use two types of approximation to obtain the distribution of the statistics of interest. The first relies on the asymptotic distribution of the estimator of the restriction in the quadratic form and uses the parametric bootstrap to approximate the asymptotic distribution of the Wald-type statistic. The second applies the residual bootstrap to obtain the distribution of interest. The residual bootstrap delivers critical values, which control empirical size and show good empirical power even in small samples. The critical values obtained by parametric bootstrap show better power properties, yet struggle to deliver correct size in small samples. The empirical size gets closer to the nominal one, however, when the sample size increases. The bootstrap approximations do not provide asymptotic refinements because the test statistic contains nuisance

parameters.

The contributions of this paper could be summarized as follows. First, it suggests a new approach to handle potential singularity in the covariance matrix of the Wald test, whereby it is shown that the rank condition of Andrews (1987) is not necessary for the approach to be valid. Second, we apply the result of Chen and Huan (1997) and Freidlin (1968) to an econometric context to show that the square root of a matrix is a continuously differentiable function of the elements of the matrix. Third, continuity of the square root of the matrix permits to the use of the continuous mapping theorem to claim that the sample version of the square root of the matrix converges to its population counterpart. This allows us to approximate the asymptotic distribution of the test statistics by the parametric bootstrap. Fourth, the distribution of the test statistic could be approximated by other bootstrap techniques, for example by the residual bootstrap. In an application involving multi-step causality testing considered in Lütkepohl and Burda (1997), the proposed approach demonstrates good small sample size properties. We compare the asymptotic versions of the tests with one another and to the residual bootstrap versions as well. It turns out that empirical size is controlled well for the critical values obtained by the residual bootstrap. Moreover, for some of the considered data generating processes, the proposed test has better power than the alternative tests presented in the literature.

The remainder of this paper is organized as follows. We start by presenting the testing procedure in a general setup in Section 2. In the next section we discuss the results of Chen and Huan (1997) and Freidlin (1968) regarding the square root of a symmetric, positive semi-definite matrix that are later used to approximate the distribution of of proposed test statistic by the parametric and residual bootstrap in Section 4. In Section 5 a simulation study is performed to mimic to the multi-step non-causality problem considered in Lütkepohl and Burda (1997). The focus here is to learn about the relative merits of the different procedures under consideration. Section 6 elaborates on the results obtained in the simulation study and considers an example illustrating how regularization techniques change the information in the data. Section 7 provides a conclusion. The proofs of all theoretical results, including the validity of the proposed approach, are presented in the appendix.

2 Testing procedure

Let $\boldsymbol{\beta}$ be a $s \times 1$ vector of the parameters of interest. One is interested in testing the following null hypothesis

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{R}(\boldsymbol{\beta}) = \boldsymbol{c}, \\ \mathrm{H}_A &: \boldsymbol{R}(\boldsymbol{\beta}) \neq \boldsymbol{c}, \end{aligned}$$

where $\boldsymbol{R} : \mathbb{R}^k \to \mathbb{R}^r$ is a continuously differentiable function and $\boldsymbol{c}$ is a $r \times 1$ matrix of known constants.

Suppose that regularity conditions hold such that there is an is asymptotically normal esti-

mator of $\boldsymbol{\beta}$ available:

$$\sqrt{T}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}_s, \boldsymbol{\Sigma}),$$

where T is the sample size. Then, the usual Wald test statistic is

$$W = T\,(\boldsymbol{R}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{c}) \left(\frac{\partial \boldsymbol{R}(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}'}\, \widehat{\boldsymbol{\Sigma}}\, \frac{\partial \boldsymbol{R}(\widehat{\boldsymbol{\beta}})'}{\partial \widehat{\boldsymbol{\beta}}} \right)^{-1} (\boldsymbol{R}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{c})',$$

in which the inverse of the weighting matrix exists. Also, $\frac{\partial \boldsymbol{R}(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}'}$, $\boldsymbol{R}(\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\Sigma}}$ are consistent estimators of $\frac{\partial \boldsymbol{R}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$, $\boldsymbol{R}(\boldsymbol{\beta})$ and $\boldsymbol{\Sigma}$, respectively. If H_0 is true and

$$\boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\beta})} = \frac{\partial \boldsymbol{R}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\, \boldsymbol{\Sigma}\, \frac{\partial \boldsymbol{R}(\boldsymbol{\beta})'}{\partial \boldsymbol{\beta}}$$

is non-singular, $W \overset{a}{\sim} \chi^2(r)$. Things are more complicated if $\boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\beta})}$ may be singular. In this case, we suggest to consider another statistic instead of W. Denote a version of W with the identity weighting matrix by W_I,

$$W_I = T\,(\boldsymbol{R}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{c})\,(\boldsymbol{R}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{c})'. \tag{1}$$

The asymptotic distribution of W_I depends on nuisance parameters and, therefore, cumbersome. But the computation of W_I avoids the use of a problematic element – the inverse of the covariance matrix. If the matrix of the second moments may cause the problems, avoiding to use it might be a solution. It turns out that the statistic W_I is continuously differentiable with respect to the elements of $\boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\beta})}$ and does not rely on the rank condition of Andrews (1987).

3 Square-root of a matrix

In order to bootstrap the distribution of W_I, we show that the square-root of a covariance matrix of the sample converges in probability to its population counterpart. This result is based on the findings of Chen and Huan (1997) and Freidlin (1968). One of the implications of the results presented in these papers is that the square-root of a matrix is continuously differentiable with respect to the entries of the matrix. Continuity permits application of the continuous mapping theorem to derive distributional results. We start with non-random matrices in the following theorem and then consider random matrices.

Proposition 1 (Continuity of eigenvalues and a square-root of a positive semi-definite matrix) *Let $\boldsymbol{\Sigma}_T$ be a $q \times q$ real positive semi-definite matrix with eigenvalues $\lambda_1(\boldsymbol{\Sigma}_T) \geq \lambda_2(\boldsymbol{\Sigma}_T) \geq \cdots \geq \lambda_q(\boldsymbol{\Sigma}_T)$. Let $\boldsymbol{\Sigma}_T^{1/2}$ be a square-root of the matrix $\boldsymbol{\Sigma}_T$. If $\boldsymbol{\Sigma}_T \to \boldsymbol{\Sigma}$ as $T \to \infty$, then*

- $\lambda_k(\boldsymbol{\Sigma}_T) \to \lambda_k(\boldsymbol{\Sigma})$, *for all* $k = 1, \ldots, s$, *and*
- $\boldsymbol{\Sigma}_T^{1/2} \to \boldsymbol{\Sigma}^{1/2}$.

Proof. see Appendix B ■

Aforementioned proposition states that the eigenvalues and the square-root are continuous functions in the elements of the matrix. However, this result can be extended to random matrices, to be applied in an econometric context as it is done in the following theorem. We present the result for convergence in probability, emphasising, though, that a similar result holds for the almost sure convergence.

Proposition 2 (Convergence in probability) *Let* $\boldsymbol{\Sigma}_T$ *be a* $q \times q$ *real positive semi-definite matrix with eigenvalues* $\lambda_1(\boldsymbol{\Sigma}_T) \geq \lambda_2(\boldsymbol{\Sigma}_T) \geq \cdots \geq \lambda_q(\boldsymbol{\Sigma}_T)$. *Let* $\boldsymbol{\Sigma}_T^{1/2}$ *be a square-root of the matrix* $\boldsymbol{\Sigma}_T$. *If* $\boldsymbol{\Sigma}_T \xrightarrow{p} \boldsymbol{\Sigma}$ *as* $T \to \infty$, *then*

- $\lambda_k(\boldsymbol{\Sigma}_T) \xrightarrow{p} \lambda_k(\boldsymbol{\Sigma}_T)$, *for all* $k = 1, \ldots, s$, *and*
- $\boldsymbol{\Sigma}_T^{1/2} \xrightarrow{p} \boldsymbol{\Sigma}^{1/2}$.

Proof. see Appendix B ■

Remark 1 *The results regarding the convergence of the square-root of the matrix rely on the fact that the integration contour includes all positive eigenvalues of* $\boldsymbol{\Sigma}$ *asymptotically. For instance, if the contour intersects the real plane in the point* $\boldsymbol{c}$, *and* $\boldsymbol{c}$ *is* $O_p(T^{-1/3})$, *eventually all non-zero eigenvalues will be included in the contour. In light of the last sentence, the parametric bootstrap approximation is similar to the spectral cut-off of Lütkepohl and Burda (1997). However, the approximation of* W_I *is more stable, in the sense that the critical values obtained by the parametric bootstrap are less sensitive to whether the smallest eigenvalues fall below the threshold or not. Instead of reducing the number of degrees of freedom of the* χ^2 *distribution, the approximation modifies the covariance matrix* $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{R}(\boldsymbol{\beta})}$ *in our case.*

Remark 2 *In case the contour is properly selected to be sample size dependent, the contour intersects the real plane at the point* $\boldsymbol{c}$, *say,* $\boldsymbol{c}$ *being* $O_p(T^{-1/3})$, *the extended continuous mapping theorem (see, e.g. Van der Vaart, 2000, Theorem 18.11) is applicable. Thus, the results listed in the previous lemmas hold even with a sample size dependent threshold, provided it converges to zero faster than the rate of consistency of the estimator of the eigenvalues.*

Remark 3 *If* $f(x)$ *and* $f(x)\,g(x)$ *are continuous, it does not imply that* $g(x)$ *is. Consider* $f(x) = x$, $g(x) = 1/x$ *as a counterexample. This example illustrates why continuity of the normalized eigenvectors does not follow from continuity of the matrix* $\boldsymbol{A}$ *and its eigenvalues.*

Remark 4 *The square-root of a matrix is a continuously differentiable transformation in the elements of the original matrix, hence it is an appropriate pivot in the sense of Beran and Srivastava (1985). This fact, howerer, is not essential for the bootstrap approximations of the distribution of* W_I *to be valid.*

4 Approximations of the distribution

Given the results in the previous section, there are several ways to obtain the distribution of W_I. The following scheme presents two possible applications of the bootstrap, which we discuss in more details in this section.

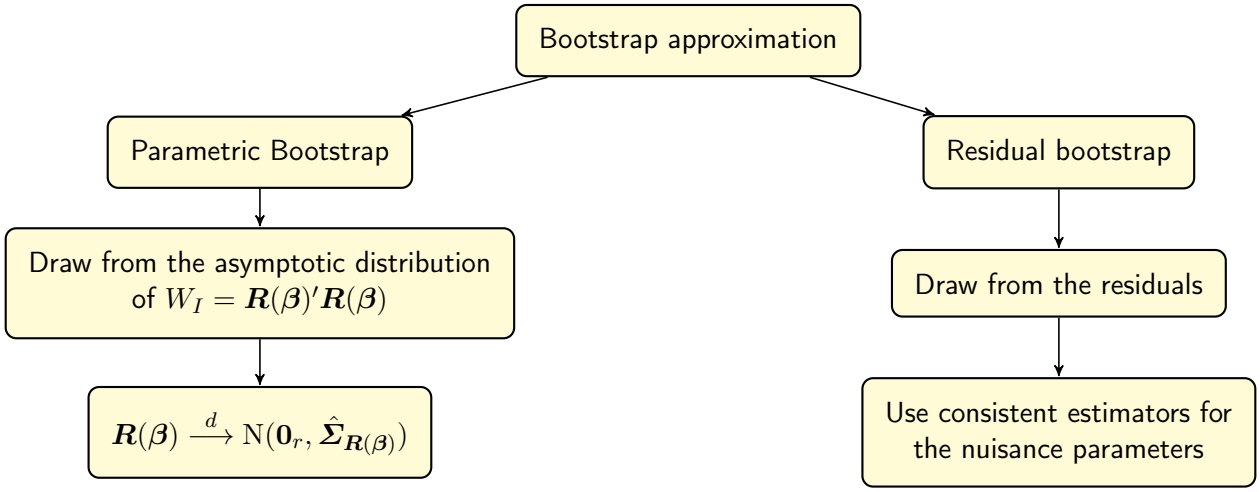

Figure 1: Two possibilities to approximate the asymptotic distribution of W_I.

One of the ways to get the distribution of W_I is to apply the parametric bootstrap based on the asymptotic approximation of the distribution of $\sqrt{T}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}_s, \boldsymbol{\Sigma})$. This approach is a particular case of the Monte Carlo tests based on consistent point estimates of the nuisance parameters of Dufour (2006) and is summarized as follows:

1. Substitute the unknown quantities in $\boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\beta})}$ by the respective estimates.
2. Draw a $r \times 1$ vector $\boldsymbol{\gamma}_l$ from the multivariate normal distribution $\mathrm{N}(\mathbf{0}_r, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{R}(\boldsymbol{\beta})})$.
3. Replicate the experiment $N_r = 10000$ times to obtain a sample of vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{10000}$.
4. For every l calculate $\omega_l = \boldsymbol{\gamma}_l'\boldsymbol{\gamma}_l$ to form a sample $\omega_1, \ldots, \omega_{10000}$.
5. Sort the resulting sample in ascended order and take $\alpha \times 10000$ element as the $\alpha - \%$ critical value.

The $\boldsymbol{\gamma}_l$, obtained in the aforementioned way, would replicate the asymptotic distribution of $\sqrt{T}\,(\hat{\boldsymbol{R}(\boldsymbol{\beta})} - \boldsymbol{R}(\boldsymbol{\beta}))$, provided we have a consistent estimator of the respective covariance matrix. Therefore, ω_l would replicate the distribution of W_I.

Another way to obtain the distribution of W_I is to bootstrap *the data*. Note that the statistic we study is not asymptotically pivotal, so the bootstrap does not provide asymptotic refinements in this case. The bootstrap may, however, provide a better approximation of the finite sample distribution of W_I. Depending on the dynamic structure of the data and the assumptions regarding the distribution of the error term, one could use different versions of the bootstrap. We apply the residual bootstrap here.

The following example illustrates this approach with an application to the statistic of interest, W_I:

1. Use the data to compute $\hat{\beta}$.

2. Generate a bootstrap sample of size T by sampling the distribution corresponding to $\hat{F}$. $\hat{F}$ could be the empirical distribution function (EDF) of the data, for which the bootstrap sample can be obtained by sampling the data randomly with replacement. If $\hat{F}$ is parametric, and hence $\widehat{F}(\cdot) = \Xi(\cdot, \widehat{\beta})$ for some function Ξ, one could sample the distribution with CDF $\Xi(\cdot, \widehat{\beta})$ to generate the bootstrap sample.

3. Compute the estimators of β and σ from the bootstrap sample. Denote the results $\widehat{\beta}^*$ and $\widehat{\sigma}^*$. Calculate the bootstrap version of W_I denoted by W_I^*.

4. Repeat the previous two steps many times to compute the empirical distribution of $\widehat{W_I}^*$. Set the critical value of interest, $z^*_{T,\alpha/2}$ equal to the $1-\alpha$ quantile of this distribution.

In the appendix we prove that both of the procedures described above are consistent.

Next we discuss how to generate the bootstrap data satisfying the null. Suppose the null hypothesis concerns only a subset of the parameters of the model – e.g., a bivariate VAR(1), and the null hypothesis concerns only one parameter of the coefficient matrix. Then, the null hypothesis makes an explicit restriction on this specific parameter in the coefficient matrix, whereas all the other coefficients can be freely chosen, yet the data would satisfy the null. In general, there is no clear answer to the question whether one should generate the bootstrap data under the null using a constrained estimator or modify the null when constructing the test statistic for the bootstrap data (see, e.g MacKinnon, 2006 for the details). There are several supporting arguments for the former choice. Firstly, the bootstrap data satisfy the null that we test for the real data. Secondly, the constrained estimator is more efficient under the null for the estimation of the nuisance parameters. The alternative way is to use the unconstrained estimator and modify the null hypothesis for the bootstrap sample such that the null hypothesis is in accordance with the bootstrap data. The advantage of this consistent testing procedure is that the covariance matrix of the restriction imposed for the bootstrap data coincides with the covariance matrix of the true data. [3]

The next section deals with a multivariate VAR process when the testing procedure is applied to multi-step non-causality testing. If the past of the variable y helps to predict the variable

[3]Our findings for the DGP considered in the following section show that the latter procedure yields better power.

x one period ahead (conditional on the past of the additional set of variables comprised in the vector $\boldsymbol{z}$) the variable y is Granger-causes of x. If the causal link works across several time periods or intermediate variables, Dufour and Renault (1998) argue that the aforementioned definition of causality suggested by Granger (1969) may be too restrictive. In these cases it is natural to generalize the original causality concept to multi-step causality defined as the potential of the variable y to predict x h step ahead.

5 Simulation study

Let us consider the DGP from Lütkepohl and Burda (1997).

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \boldsymbol{u}_t, \qquad t = 1, \ldots, T. \tag{2}$$

where $\boldsymbol{y}_t$ a is vector of 3 time series denoted by $\boldsymbol{y}_t = (x_t, y_t, z_t)'$ and $\boldsymbol{u}_t \sim \mathrm{N}(\boldsymbol{0}_3, \boldsymbol{\Sigma})$. Note that

$$\boldsymbol{A}_1 = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix}.$$

Let $\boldsymbol{\alpha} = \mathrm{vec}(\boldsymbol{A}_1)$.

The goal is to test the null hypothesis of multi-step Granger non-causality running from y_t to x_t, i.e. $\mathrm{H}_0 : y_t \overset{(\infty)}{\nrightarrow} x_t$. As shown in Lütkepohl and Burda (1997), one of the ways to do that is to test two restrictions on $\boldsymbol{\alpha}$ of the following form

$$\mathrm{H}_0 : \boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{0},$$
$$\mathrm{H}_a : \boldsymbol{R}(\boldsymbol{\alpha}) \neq \boldsymbol{0},$$

where

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{bmatrix} \alpha_{xy} \\ \alpha_{xx}\alpha_{xy} + \alpha_{xy}\alpha_{yy} + \alpha_{xz}\alpha_{zy} \end{bmatrix}. \tag{3}$$

There are three parameter settings satisfying these restrictions:

$$\alpha_{xy} = \alpha_{xz} = 0, \alpha_{zy} \neq 0, \tag{4}$$
$$\alpha_{xy} = \alpha_{zy} = 0, \alpha_{xz} \neq 0, \tag{5}$$
$$\alpha_{xy} = \alpha_{xz} = \alpha_{zy} = 0. \tag{6}$$

Importantly for the approach concerned, the matrix of first-order partial derivatives of the function $\boldsymbol{R}(\boldsymbol{\alpha})$, given by

$$\frac{\partial \boldsymbol{R}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \alpha_{xy} & 0 & 0 & \alpha_{xx}+\alpha_{yy} & \alpha_{xy} & \alpha_{zy} & 0 & 0 & \end{bmatrix},$$

could be singular if (6) holds. In this case, the standard Wald statistic does not have an asymptotic $\chi^2(2)$ distribution.

To test the null of multi-step non-casuality, $\mathrm{H}_0 : \boldsymbol{R}(\boldsymbol{\alpha}) = \mathbf{0}$, we consider the standard Wald test statistic, three different modified versions of the Wald statistic together with the approach we suggest in this paper. The objective of the following simulation study is twofold: First, to investigate the small sample properties of the proposed method. Second, to compare it to the regularization based modifications of Wald test available in the literature.

Throughout this section data are generated according to (2), using an initialization period of length $B = 100$. As in Lütkepohl and Burda (1997) the coefficient matrix $\boldsymbol{A}_1$ is selected to be upper triangular, i.e. $\alpha_{yx} = \alpha_{zx} = \alpha_{zy} = 0$. Hence, we have that $x_t \not\to y_t$ and $x_t \not\to z_t$ and $y_t \not\to z_t$. $\alpha_{zy} = 0$ is fixed such that the causal link from y_t to x_t is determined by the coefficient α_{xy} only, i.e., the restrictions in (3) are met if $\alpha_{xy} = 0$. For the remaining, potentially non-zero, VAR coefficients we consider

$$\begin{aligned} &\alpha_{xx} = \alpha_{yy} = \alpha_{zz} = \theta, \theta \in \{-0.99, -0.9, -0.3, 0.3, 0.9, 0.99\}, \\ &\alpha_{yz} = 0.5, \\ &\alpha_{xz} \in \{0, 0.5\}. \end{aligned}$$

The values of the diagonal elements of $\boldsymbol{A}_1$ vary in order to cover the possibility of being well inside the stationary region as well as close to the non-stationary region. Stability of the system is guaranteed under both the null and the alternative hypothesis. We assume the the true lag or the system is known when the model is estimated and the residual bootstrap is applied. All simulations are carried out with $MATLAB$ 7.12.0

In the tables below, W denotes the standard Wald statistic, W_{LB}^m is the modified Wald statistic and W_{LB} is the spectral cut-off regularized Wald statistic of Lütkepohl and Burda (1997), W_{DF4} is the regularized Wald statistic based on a super-consistent estimator of the eigenvalues at the threshold c of Dufour et al. (2011), and W_I and W_{Ib} correspond to the statistic that we suggest. The first uses the critical values obtained by parametric bootstrap, whereas the second one relies on residual bootstrap for this purpose. Note that, as suggested by the authors, W_{LB}^m is calculated with $\lambda = 0.1$, λ is the parameter that determines the variance of the added noise component. W_{LB} is computed with the theoretically preferred value $c_1 = \hat{\lambda}_1 T^{-1/3}$, where $\hat{\lambda}_1$ is the estimate of the largest eigenvalue of the covariance matrix of the restrictions. W_{DF4} is computed with the threshold 0.1 and the Variance Regularization Function defined by equation (7.3) in Dufour et al. (2011). Pseudo-standard normal random numbers are generated for the, ϵ_t and to calculate W_{LB}^m.

5.1 Behaviour under the null

In the right-hand panels of Tables 1 and 2 α_{xz} is set to 0.5 and all statistics are expected to be appropriate. Whereas, in the left-hand panels of both tables α_{xz} equals 0 and the standard Wald statistic fails to have its usual limiting $\chi^2(2)$ distribution under the null. Therefore, the relative rejection frequencies of the standard Wald statistic in the left-hand panel are much lower than 5%, even for $T = 1000$. W_{DF4} is also undersized, which is in line with the claim that this statistic is conservative made in Dufour et al. (2011), although for $T = 100$ the statistic is oversized for some of the cases. Size distortions of the other test statistics depend on the sample size and persistency of the data as Table 1 illustrates. Moreover, W_{Ib} is the only test, which controls size uniformly. That is to be expected since it is the only test using critical values computed by the residual bootstrap. Note that the standard error of a 5 % rejection probability estimated from 5000 independent replications is $\sqrt{0.05 \times 0.95/5000} \approx 0.0031$.

Table 1: Relative rejection frequencies of Wald, modified Wald tests and W_I under the null

$T = 100$	$\alpha_{xz} = 0$						$\alpha_{xz} = 0.5$					
α_{ii}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}
-0.99	24.3	17.2	22.6	0.0	22.6	2.4	23.7	20.7	18.1	15.2	18.2	4.9
-0.9	5.5	7.4	9.8	6.9	9.7	4.3	8.6	8.0	7.7	4.3	7.8	4.0
-0.3	1.9	3.6	5.8	1.8	5.7	5.0	4.7	4.8	5.6	9.3	5.4	4.6
0.3	1.9	3.8	5.9	1.9	5.7	5.0	4.4	4.9	5.9	9.6	5.7	4.9
0.9	7.2	9.3	11.7	10.1	11.7	3.7	13.8	12.9	12.7	6.8	12.8	5.7
0.99	26.4	22.1	28.6	0.5	28.6	2.5	36.2	32.6	33.1	24.8	33.2	9.5

Note: Reported sample size is 100. The left hand side corresponds to the irregular case, whereas the right hand side contains the results for the regular case.

Table 2: Relative rejection frequencies of Wald, modified Wald tests and W_I under the null

$T = 1000$	$\alpha_{xz} = 0$						$\alpha_{xz} = 0.5$					
α_{ii}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}
-0.99	5.0	6.6	8.1	0.0	8.0	2.3	9.4	8.4	7.6	3.2	7.6	1.9
-0.9	1.8	5.1	5.6	1.8	5.5	4.6	5.7	5.6	6.1	1.9	6.1	4.9
-0.3	1.5	4.6	4.9	1.5	4.9	4.7	4.5	4.8	4.4	7.2	4.6	4.7
0.3	1.2	4.4	5.0	1.1	5.0	5.0	4.4	4.8	4.5	7.2	4.7	4.8
0.9	2.1	5.7	5.5	2.3	5.5	4.5	6.1	5.8	6.0	2.1	5.9	4.3
0.99	7.7	10.2	12.3	0.0	12.3	2.1	13.9	12.1	12.6	7.5	12.7	2.6

Note: Reported sample size is 1000. The left hand side corresponds to the irregular case, whereas the right hand side contains the results for the regular case.

5.2 Behaviour under the alternative

Next, we examine the power properties of the tests. The data generated under the alternative violates the null in the following way

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{bmatrix} \alpha_{xy} \\ \alpha_{xx}\alpha_{xy} + \alpha_{xy}\alpha_{yy} + \alpha_{xz}\alpha_{zy} \end{bmatrix} = \begin{bmatrix} \delta \\ (\alpha_{xx} + \alpha_{yy})\delta \end{bmatrix},$$

where δ determines how far the data is from the null. We consider the fixed alternatives, following the literature.

In both panels of Tables 3 and 4, all statistics seem to incur with power losses if the data is well inside in the stationary region. The power properties improve when the sample size grows as well as when the generated data is more persistent. Note that the relative merit of the tests is stable across different sample sizes and values of α_{ii}. W_{LB}, W_{DF4} and W_I are more powerful than W_{LB}^m for this specific DGP. The test that uses the critical values calculated by the residual bootstrap, W_{Ib}, lacks power in some cases compared to its competitors. This is to be expected given the results under the null. One of the possible ways to take into account the size distortions of some of the tests is to consider size adjusted power. As shown in Table 5, the power loss for W_{Ib}, compared to other tests, is smaller if we use the size adjusted power to compare the tests.

Table 3: Relative rejection frequencies of Wald, modified Wald tests and W_I under the alternative

$T = 100$	$\alpha_{xz} = 0$						$\alpha_{xz} = 0.5$					
α_{ii}	W	W_{LB}^m	W_{LB}	W_{DF4}	W_I	W_{Ib}	W	W_{LB}^m	W_{LB}	W_{DF4}	W_I	W_{Ib}
-0.99	99.9	99.9	99.9	99.9	99.9	99.9	100	100	100	100	100	99.9
-0.9	80.3	76.4	84.8	85.6	84.8	72.4	69.2	68.5	73.0	73.9	73.0	54.1
-0.3	5.1	7.4	12.3	4.8	11.9	10.7	8.5	8.8	11.0	13.8	11.1	9.8
0.3	5.7	8.4	13.4	5.4	13.1	11.7	10.0	10.0	12.5	16.0	13.3	11.6
0.9	81.6	79.5	87.1	86.2	87.0	69.8	49.4	47.5	54.8	39.2	54.8	35.9
0.99	99.3	99.3	99.5	99.5	99.5	97.6	99.0	98.8	98.5	98.4	98.6	83.5

Note: Reported sample size is 100. The left hand side corresponds to the irregular case, whereas the right hand side contains the results for the regular case. $\delta = 0.0632$.

Table 4: Relative rejection frequencies of Wald, modified Wald tests and W_I under the alternative

$T = 1000$	$\alpha_{xz} = 0$						$\alpha_{xz} = 0.5$					
α_{ii}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}
-0.99	100	100	100	100	100	100	100	100	100	100	100	100
-0.9	100	100	100	100	100	100	100	100	100	100	100	100
-0.3	48.3	55.7	66.2	47.8	66.2	66.3	53.9	54.1	54.9	56.5	62.1	61.9
0.3	48.4	56.0	67.5	47.8	67.4	67.2	54.2	54.0	54.6	57.4	62.6	62.2
0.9	100	100	100	100	100	100	100	100	100	100	99.9	100
0.99	100	100	100	100	100	100	100	100	100	100	100	100

Note: Reported sample size is 1000. The left hand side corresponds to the irregular case, whereas the right hand side contains the results for the regular case. $\delta = 0.0632$.

Table 5: Relative rejection frequencies of Wald, modified Wald tests and W_I under the alternative

$\alpha_{ii} = 0.9$	$\alpha_{xz} = 0$						$\alpha_{xz} = 0.5$					
T	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}
100	77.9	70.2	74.6	76.6	74.7	70.6	42	35.9	36.8	20.2	36.8	36.1
1000	100	100	100	100	100	100	100	100	100	100	100	100

Note: Reported sample sizes are 100 and 1000. The left hand side corresponds to the irregular case, whereas the right hand side contains the results for the regular case. $\delta = 0.0632$.

5.3 Behaviour under the alternative when only the second restriction is violated

Let us check the power properties of the tests when only the second restriction is violated. It is important to explore this case to see how the modifications of the Wald test we consider handle the situation, in which the two restrictions suggest discordant evidence regarding the violation of the null. As in Lütkepohl and Burda (1997), we generate 5000 realizations of the VAR(1) process of sample sizes 100 and 1000 with the coefficient matrix

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.3 & 0 & \alpha_{xz} \\ 0.7 & 0.3 & 0.25 \\ 0.5 & 0.4 & 0.3 \end{pmatrix},$$

with $\alpha_{xz} = \delta$ and $\delta = \{0.0632, 0.1264\}$. Stationarity of the process is guaranteed for both values of δ. Notice that this case is favorable to the standard Wald test compared to the modified statistics, among which W_{DF4} and W^m_{LB} perform best with W_{LB} and W_I having comparable power properties.

Table 6: Relative rejection frequencies of Wald, modified Wald tests and W_I

T=100	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}
$\delta = 0$	4.8	5.8	6.2	4.4	6.2	5.2
$\delta = 0.0632$	7.8	7.2	6.2	6.9	6.1	5.2
$\delta = 0.1264$	22.8	15.8	7.3	6.9	6.9	5.6

T=1000	W	W^m_{LB}	W_{LB}	W_{DF4}	W_I	W_{Ib}
$\delta = 0$	5.3	5.4	5.1	2.6	5.1	5.0
$\delta = 0.0632$	75.9	46.8	15.9	53.2	10.4	10.1
$\delta = 0.1264$	100	97.6	33.4	99.8	51.0	49.8

Note: Reported sample sizes are 100 and 1000.

5.4 Residual bootstrap versions of the tests

In the last part of the simulation study, we use the residual based bootstrap to calculate the critical values for all the test statistics in order to check how sensitive the results of the testing are depending on whether the asymptotic or bootstrap critical values are used. As becomes evident from the previous results, W_{Ib} controls size better than W_I and all the other alternatives. Thus, we expect that the bootstrap critical values for other tests would control size as well, which turns out to be the case. In the next set of tables, we present the empirical rejection frequencies of different tests under the alternative. Table 7 displays the relative rejection frequencies of the tests using asymptotical and bootstrap critical values.

Table 7: Relative rejection frequencies of Wald, modified Wald tests and W_I

	W		W^m_{LB}		W_{LB}		W_{DF4}		W_I	
T	100	1000	100	1000	100	1000	100	1000	100	1000
Asymptotic	5.7	48.4	8.4	56.0	13.4	67.5	10.0	67.1	13.1	67.4
Bootstrap	2.8	38.7	7.2	56.1	10.8	66.9	5.8	66.7	11.7	67.2

Note: Reported sample sizes are 100 and 1000. $\alpha_{ii} = 0.3$, irregular setup. The first row corresponds to asymptotic critical values, whereas the second to bootstrap critical values. $\delta = 0.0632$.

The results in Tables 8, 9 and 10, presenting a similar exercise for various DGPs under the alternative, suggest that the relative performance of W_{LB}, W^m_{LB}, W_{DF4} and W_I depends on the DGP. Overall, the bootstrap versions of the test control size well for all alternatives. W_I and its bootstrap counterpart work better than the others if the data is well inside the stationary region ($\mid \alpha_{ii} \mid = 0.3$), performing worse than the others if the data is approaching the non-stationary region ($\mid \alpha_{ii} \mid = 0.9$). The tests demonstrate similar behavior if the data is very close to non-stationary region ($\mid \alpha_{ii} \mid = 0.99$).

The explanation of these results may be related to the fact that the test suggested in this paper does not depend on tuning parameters, whereas all the other modifications of the standard Wald test do. Moreover, regularization modifies the information in the data, whereas the proposed test does not.

Note that, we consider only one application to compare the relative merits of the tests. An independent research performed in Duchesne and Francq (2014) considers is in line with our findings. In practice, one has to come up with a choice of tuning parameters for all the tests, except the one proposed in this paper. Firstly, this is a clear advantage of the procedure we suggest, since there is no general guidelines regarding the choice of tuning parameters. This choice also affects the relative performance of the tests in different circumstances. In the next section we illustrate the points considered in the previous paragraph by studying the asymptotic power properties for a very simple DGP.

6 Regularize or not?

Suppose the true rank of the restriction in the model considered in the previous section is 1. The test we suggest would use the estimate of the population covariance matrix of the restriction, which has an incorrect rank with probability one. Nevertheless, the test would use a "close" approximation of the asymptotic covariance matrix of the restriction. The approximation is close in the sense that this estimator converges in probability to the true value and that, if one uses the parametric bootstrap to approximate the asymptotic distribution of the test, the difference between the estimated square root of the matrix and the true square root of the matrix in terms of a given matrix norm is bounded and can be bounded with arbitrary precision.

All the other modifications of the standard Wald test are regularization approaches, so they modify the information contained in the data. They either substitute the small eigenvalues with zeros (W_{LB}) or add some noise to the estimate of the restriction (W_{LB}^m) to make it invertible or substitute the small eigenvalue by some threshold value (W_{DF4}). Hence, depending on the true rank of the system and the way the data violates the null hypothesis, this modification may be advantageous or disadvantageous for the regularization approaches. For example, in case $\alpha_{ii} = 0.3$ and the regular restriction, the simulations suggest that, in many cases the second restriction is omitted by W_{LB} because the second eigenvalue falls below the threshold. This implies that respective information is not used. It turns out to be disadvantageous for W_{LB} in this case, It may, however, be the case that dropping a relevant restriction increases power, namely, if the violation of the null is not that prominent for this restriction compared to the other one. Clearly, if a regularization approach drops a restriction – the other number for the degrees of freedom is used to determine the appropriate critical value – it may be that such a change causes higher power. Note that whether such a change actually occurs, depends on the data, and on the specific regularization approach that modifies the covariance matrix. In some sense, regularization always adds additional information, and researchers should be aware of this.

In this section we consider a simple model to explain the results of the simulation study

more formally. The model is similar to the one studied in Duchesne and Francq (2014) and can be summarized, as follows:

$$\boldsymbol{y}_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} + \begin{pmatrix} u_{x,t} \\ u_{y,t} \end{pmatrix}, \qquad t = 1, \ldots, T. \tag{7}$$

where $\boldsymbol{y}_t$ is a vector of $s = 2$ time series denoted by $\boldsymbol{y}_t = (x_t, y_t)'$ and $\boldsymbol{u}_t \sim \mathrm{N}(\boldsymbol{0}_s, \boldsymbol{\Sigma})$.

We test whether the mean of the series x_t and y_t, is in line with some hypothesized values. It could be done testing two restrictions on $(\mu_x, \mu_y)'$ as follows

$$\mathrm{H}_0 : \boldsymbol{R}(\boldsymbol{\mu}) = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \begin{pmatrix} \mu_{0,x} \\ \mu_{0,y} \end{pmatrix},$$
$$\mathrm{H}_a : \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \neq \begin{pmatrix} \mu_{0,x} \\ \mu_{0,y} \end{pmatrix},$$

where $(\mu_{0,x}, \mu_{0,y})'$ are some hypothesized values for the means of the processes. Let us assume $\boldsymbol{\Sigma}$ is diagonal

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix} = \boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\mu})},$$

so the asymptotic distributions of the tests is fairly simple. The goal of this section is to derive and compare the power functions for various tests considered in this paper. It is done for a local alternative. W_{LB}^m follows an asymptotic non-central χ^2-distribution,

$$W_{LB}^m \overset{a}{\sim} \chi^2(2, ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))' (\boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\mu})})^{-1} ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))'),$$

as discussed in Lütkepohl and Burda (1997), whereas the asymptotic distribution of W_{LB} depends on the rank of $\boldsymbol{\Sigma}$, j_1:

$$W_{LB} \overset{a}{\sim} \chi^2(j_1, ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))' \boldsymbol{V} \boldsymbol{\Lambda}_c \boldsymbol{V}' ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))'),$$

where $\boldsymbol{V}$ is determined by the Cholesky decomposition of matrix $\boldsymbol{\Sigma}$, $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$, and $\boldsymbol{\Lambda}_c$ is a matrix that is obtained from $\boldsymbol{\Lambda}$ by inverting the eigenvalues that are above the threshold c and substituting by zeros the eigenvalues that are below it.

Likewise, the asymptotic distribution of W_{DF} is

$$W_{DF} \overset{a}{\sim} \begin{array}{l} \left((\hat{\mu}_x, \hat{\mu}_y) - (\mu_x, \mu_y)\right)' \boldsymbol{\Sigma}^R(c) ((\hat{\mu}_x, \hat{\mu}_y) - (\mu_x, \mu_y)) + 2 ((\hat{\mu}_x, \hat{\mu}_y) - (\mu_x, \mu_y))' \boldsymbol{\Sigma}^R(c) ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y})) + \\ + ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))' \boldsymbol{\Sigma}^R(c) ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y})), \end{array}$$

where $\boldsymbol{\Sigma}^R(c)$ is the regularized inverse with a fixed threshold c.

Next, the asymptotic distribution of W_I can be described as

$$W_I \overset{a}{\sim} \sigma_x \chi_x^2(1) + \sigma_y \chi_y^2(1) ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))' \boldsymbol{\Sigma} ((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y}))',$$

where the subscripts x and y for the χ^2 random variables are added to make clear that the two are different. Finally, the standard Wald test has the non-central χ^2 distribution,

$$W \overset{a}{\sim} \chi^2 \left(2, \left((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y})\right)' \boldsymbol{\Sigma}^{-1} \left((\mu_x, \mu_y) - (\mu_{0,x}, \mu_{0,y})\right)'\right),$$

Consider that $\boldsymbol{\Sigma}$ is fixed to the following matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 0.05 \end{pmatrix} = \boldsymbol{\Sigma}_{\boldsymbol{R}(\boldsymbol{\mu})},$$

such that σ_y is selected below the threshold value $c = 0.1$. So, depending on the exact DGP under the alternative, the performance of the various tests differs in terms of power.

$$\begin{aligned} &Case\ 1 && \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \begin{pmatrix} \mu_{0,x} + \delta \\ \mu_{0,y} + \delta \end{pmatrix}, \\ &Case\ 2 && \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \begin{pmatrix} \mu_{0,x} + \delta \\ \mu_{0,y} \end{pmatrix}, \\ &Case\ 3 && \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \begin{pmatrix} \mu_{0,x} \\ \mu_{0,y} + \delta \end{pmatrix}, \end{aligned}$$

where $\delta = a/\sqrt{T}$.

In this model, the rank of the restriction is determined by the rank of $\boldsymbol{\Sigma}$. Thus, the standard Wald test is optimal given the rank of $\boldsymbol{\Sigma}$ is full. It does not guarantee, however, that the standard Wald test has the highest power. Clearly, some of the modified Wald tests may perform better. For example, W_{LB} drops the second restriction while calculating the value of the test statistic, which may lead to better power if the DGP is Case 2. At the same time, if the true DGP is Case 3, W_{LB} lacks power compared to the standard Wald test because it omits the restriction, which contains essential information. Therefore, W_{LB} does not have power for some alternatives. W_{DF} and W_{LB}^m, on the contrary, have power for any DGP, being less efficient in Case 3 than the standard Wald test. W_{DF} uses information in the second restriction but blow it up with a smaller value due to the fact that the inverse is regularized. In the example above, the second restriction is multiplied by 10$=1/0.1$ and not by 20 as is the case for the standard Wald test. W_{LB}^m adds some noise to the estimate of the restriction, so the non-centrality parameter is smaller than for the standard Wald test. Finally, the asymptotic power properties of W_1 coincide with the standard Wald test if the rank of $\boldsymbol{\Sigma}$ is full.

When the rank is not full, it is not clear which test statistic has the best power asymptotically since it depends on the truncation parameters of the regularized Wald tests and on whether the data is generated by Case 1, 2 or 3. Clearly, if the data is generated by Case 3, and the rank of $\boldsymbol{\Sigma}$ is not full, the statistics that treat the two restrictions as if they have different source of information should perform better. If the second restriction is exploded by a 20, as it is the case for W_{DF}, it increases the chance of the statistic to be larger than the critical value compared to W_{LB}. Also, the relative chance of W_{DF} to be larger than its critical value may be higher

than for W_1 since the former treats the two restrictions as if the originate from more or less the same source of information. If one restriction suggests that the null should not be rejected and the other one suggests the opposite, it is the magnitude of the estimation error that determines whether the null should be rejected or not. In the case of W_1, the magnitude does not affect the way one calculates the test statistic. It only affects the way the critical values are calculated. Hence, although the critical value gets smaller as the determinant of $\boldsymbol{\Sigma}$ is getting closer to zero, the critical value also depends on the magnitude of the restriction with larger variance for the example above. Therefore, small deviations from the null only in one of the restrictions do not guarantee that the power of W_1 is large. W_1 only uses information that is in the data not changing it at all, therefore, it only can tell the difference between the null and the alternative if the data suggest so.

7 Conclusions

In this paper we proposed an alternative way to handle possible singularity in the covariance matrix of the Wald test. The rank condition of Andrews (1987) is not necessary for the approach to be valid. We applied the result of Chen and Huan (1997) and Freidlin (1968) to show that the square root of the matrix is continuously differentiable in elements of the matrix. This allowed us to apply the continuous mapping theorem to claim that the sample version of the square root of the matrix converges to its population counterpart. As a result, we approximated the asymptotic distribution of the test statistics by the parametric and the residual bootstrap.

We considered the DGP from Lütkepohl and Burda (1997) in the application to multi-step Granger non-casuality in a Monte Carlo study. The results of the study showed that the parametric bootstrap approximation worked well to provide a testing procedure, which has power and size comparable to existing alternatives. When the data was well inside the stationary region, the proposed test statistic had the best power of all alternatives, whereas it lacked power compared to its competitors if the data is more persistent. Overall, most of the tests considered in this paper did not control size well especially if the data was persistent. Therefore, we used the residual bootstrap to obtain alternative critical values. The resulting approach controlled empirical size better for all the tests. When compared to the residual bootstrap versions of the other tests, the proposed test performed better when the data was well inside the stationary region and worse when the data was more persistent. Thus, the relative merits across various tests are similar for asymptotic and residual bootstrap approximations of the tests distributions. Compared to the modifications of the Wald test based on regularization techniques, the test statistic proposed in this paper does not involve arbitrary truncation parameters for which no practical guidelines are available and does not modify the information in the data.

References

Andrews, D. W. (1987). Asymptotic results for generalized wald tests. *Econometric Theory*, 3(3):348–358.

Andrews, D. W. (1988a). Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics*, 37(1):135–156.

Andrews, D. W. (1988b). Chi-square diagnostic tests for econometric models: Theory. *Econometrica: Journal of the Econometric Society*, pages 1419–1453.

Beran, R. and Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, pages 95–115.

Chen, Z. and Huan, Z. (1997). On the continuity of the mth root of a continuous nonnegative definite matrix-valued function. *Journal of Mathematical Analysis and Applications*, 209(1):60 – 66.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians: An Introduction for Econometricians*. Oxford university press.

Duchesne, P. and Francq, C. (2014). Multivariate hypothesis testing using generalized and {2}-inverses–with applications. *Statistics*, (ahead-of-print):1–22.

Dufour, J.-M. (2006). Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, 133(2):443–477.

Dufour, J.-M. and Renault, E. (1998). Short run and long run causality in time series: theory. *Econometrica*, pages 1099–1125.

Dufour, J.-M., Valéry, P., and Montréal, H. (2011). Wald-type tests when rank conditions fail: a smooth regularization approach. Technical report.

Freidlin, M. I. (1968). On factorization of a nonnegatively definite matrix. *Teoriya Veroyatnostei i ee Primeneniya*, 13(2):375–378.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton University Press.

Horn, R. A. and Johnson, C. (1985). *Matrix analysis*. Cambridge University Press.

Horowitz, J. L. (2001). Chapter 52 the bootstrap. In Heckman, J. and Leamer, E., editors, *Handbook of econometrics*, volume 5 of *Handbook of Econometrics*, pages 3159 – 3228. Elsevier.

Katō, T. (1995). *Perturbation theory for linear operators*, volume 132. Springer Verlag.

Knopp, K. (1996). The residue theorem. *Theory of Functions, Dover, New York, Part I*, pages 129–134.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Lütkepohl, H. and Burda, M. M. (1997). Modified wald tests under nonregular conditions. *Journal of Econometrics*, 78(2):315–332.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics*. *Economic Record*, 82(s1):S2–S18.

Mitra, S. K. (1980). Generalized inverse of matrices and their application to linear models. *Handbook of Statistics*, 1:471–512.

Moore, D. S. (1977). Generalized inverses, wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, 72(357):131–137.

Newey, W. K. (1987). Specification tests for distributional assumptions in the tobit model. *Journal of Econometrics*, 34(1):125–145.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

Appendix A

Table 8: Relative rejection frequencies of Wald, modified Wald tests and W_I

W		W_{LB}^m		W_{LB}		W_{DF4}		W_I		
T	100	1000	100	1000	100	1000	100	1000	100	1000
Asymptotic	10.0	54.2	10.0	54.0	12.5	54.6	7.2	48.2	13.3	62.6
Bootstrap	82.6	54.2	8.7	54.3	9.2	54.8	7.2	47.8	11.5	61.9

Note: Reported sample sizes are 100 and 1000. $\alpha_{ii} = 0.3$, regular setup. The first row corresponds to asymptotic critical values, whereas the second to bootstrap critical values. $\delta = 0.0632$.

Table 9: Relative rejection frequencies of Wald, modified Wald tests and W_I

W		W_{LB}^m		W_{LB}		W_{DF4}		W_I		
T	100	1000	100	1000	100	1000	100	1000	100	1000
Asymptotic	81.6	100	79.5	100	87.1	100	87.0	100	87.0	100
Bootstrap	75.9	100	73.2	100	78.2	100	77.8	100	70.5	100

Note: Reported sample sizes are 100 and 1000. $\alpha_{ii} = 0.9$, irregular setup. The first row corresponds to asymptotic critical values, whereas the second to bootstrap critical values. $\delta = 0.0632$.

Table 10: Relative rejection frequencies of Wald, modified Wald tests and W_I

W		W_{LB}^m		W_{LB}		W_{DF4}		W_I		
T	100	1000	100	1000	100	1000	100	1000	100	1000
Asymptotic	49,4	100	47.5	100	54.8	100	55.0	100	54.8	100
Bootstrap	36.8	100	37.1	100	43.1	100	41.8	100	36.1	100

Note: Reported sample sizes are 100 and 1000. $\alpha_{ii} = 0.9$, regular setup. The first row corresponds to asymptotic critical values, whereas the second to bootstrap critical values. $\delta = 0.0632$.

Appendix B

Asymptotic consistency of the approximations of W_I

First, we show that the square root of a matrix is a continuous transformation in the elements of the matrix, therefore prepositions 1 and 2 hold. Then, we show that CDF of W_I is a continuous function with respect to the covariance matrix of the restriction. Therefore the bootstrap approximations of CDF converge uniformly in probability to the asymptotic CDF of W_I.

7.0.1 Proof of propositions 1 and 2

To prove these propositions, we follow Chen and Huan (1997) and Freidlin (1968) and let $\boldsymbol{A}(\boldsymbol{x}) = (a^{ij}(\boldsymbol{x}))$ be a non-negative definite $n \times n$ matrix, whose elements depend on the point $\boldsymbol{x} \in \mathbb{R}^N$. A real or complex-valued function u on the N-dimensional Euclidean space satisfies a Hölder condition, when there is a non-negative real constant K such that $\mid u(\boldsymbol{x}) - u(\boldsymbol{y}) \mid \leq K \mid \boldsymbol{x} - \boldsymbol{y} \mid^{\alpha}$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in the domain of u. Then, a Hölder space is a functional space, consisting of functions satisfying a Hölder condition. Let $\alpha \in [0, 2]$ and $\mathbb{E} \subset \mathbb{R}^N$ be a non-empty open set. Define $C^0(\mathbb{E}) = C(\mathbb{E})$ as the set of all continuous functions in $\mathbb{E}$. Similarly, $C^1(\mathbb{E})$ is the subset of $C(\mathbb{E})$, every member of which has continuous first order partial derivatives in $\mathbb{E}$. For $\alpha \in (0, 1]$, $C^{\alpha}(\mathbb{E}) = C^{0,\,\alpha}(\mathbb{E})$ is the subset of $C(\mathbb{E})$ for which each member u satisfies the following: for any bounded closed subset $\mathbb{G}$ of $\mathbb{E}$ there exists a constant K such that $\mid u(\boldsymbol{x}) - u(\boldsymbol{y}) \mid \leq K \mid \boldsymbol{x} - \boldsymbol{y} \mid^{\alpha}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{G}$. Moreover, call $[u]_{\alpha,\mathbb{G}} = \inf\{K :\mid u(\boldsymbol{x}) - u(\boldsymbol{y}) \mid \leq K \mid \boldsymbol{x} - \boldsymbol{y} \mid^{\alpha}, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{G}\}$ the seminorm of u on $\mathbb{G}$ in $C^{\alpha}(\mathbb{E})$. For $\alpha \in (1, 2]$, $C^{\alpha}(\mathbb{E}) = C^{1,\,\alpha-1}(\mathbb{E})$ is the subset of $C^1(\mathbb{E})$, for which the members have first order derivatives in $C^{\alpha-1}(\mathbb{E})$. A matrix-valued function $\boldsymbol{A}(\boldsymbol{x})$ is in $C^{\alpha}(\mathbb{E})$ if all entries of $\boldsymbol{A}(\boldsymbol{x})$ are in $C^{\alpha}(\mathbb{E})$.

Let m be a finite positive integer. An $n \times n$ non-negative definite matrix $\boldsymbol{B}$ is called the m-th root of $\boldsymbol{A}$, denoted by $\boldsymbol{A}^{1/m}$, provided that $\boldsymbol{B}^m = \boldsymbol{A}$. Let $\boldsymbol{A}(\boldsymbol{x})$ be an $n \times n$ symmetric matrix-valued function defined on a subset $\mathbb{E} \subset \mathbb{R}^N$. A function $\boldsymbol{A}(\boldsymbol{x})$ is non-negative if, for any $\boldsymbol{x} \in \mathbb{E}$, $\boldsymbol{A}(\boldsymbol{x})$ is non-negative definite. A non-negative definite $n \times n$ matrix-valued function $\boldsymbol{B}(\boldsymbol{x})$ is called the m-th root of $\boldsymbol{A}(\boldsymbol{x})$, denoted by $\boldsymbol{A}^{1/m}(\boldsymbol{x})$, provided that $\boldsymbol{B}(\boldsymbol{x})^m = \boldsymbol{A}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{E}$. Note that the m-th root of a non-negative definite matrix is uniquely defined. Also, $\mathbb{R}^N$ stands for the N-dimensional Euclidean space, and $\mid \boldsymbol{x} \mid$ for the Euclidean norm of $\boldsymbol{x}$ in $\mathbb{R}^N$. Whereas $\mid \boldsymbol{A} \mid$ denotes the norm $\max_{i,j \leq n}\{\mid A_{ij} \mid\}$.

Then, Chen and Huan (1997) show that the following statement holds.

Theorem 1 *Let $\boldsymbol{A}(\boldsymbol{x})$ be a continuous non-negative definite matrix-valued function on a subset $\mathbb{G} \subset \mathbb{R}^N$. then the m-th root function $\boldsymbol{A}^{1/m}(\boldsymbol{x})$ of $\boldsymbol{A}(\boldsymbol{x})$ is also continuous on $\mathbb{G}$.*

This statement is a generalization of the result in Freidlin (1968). The proof of this theorem relies on the following lemma

Lemma 1 *Let $\boldsymbol{A}(\boldsymbol{x})$ be a positive definite matrix-valued function on a bounded domain $\mathbb{G} \subset \mathbb{R}^N$, so $W_I\, \boldsymbol{I} \leq \boldsymbol{A}(\boldsymbol{x}) \leq w_2\, \boldsymbol{I}$ for $\boldsymbol{x} \in \mathbb{G}$, where W_I and w_2 are positive constants. Then, the m-th*

root of $\boldsymbol{A}(\boldsymbol{x})$ *can be written as*

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{1}{2\pi\, i}\int_{\Gamma} f(z)dz = \frac{1}{2\pi\, i}\int_{\Gamma} \sqrt{z}(\boldsymbol{A}(\boldsymbol{x}) - z\,\boldsymbol{I})^{-1}dz, \tag{8}$$

where $i = \sqrt{-1}$ *and* Γ *is a closed contour that contains all non-zero eigenvalues of* $\boldsymbol{A}(\boldsymbol{x})$*, and is in the right complex half plane* $\{z : \mathrm{Re}\ z > 0\}$*. Moreover,* $\boldsymbol{B}(\boldsymbol{x})$ *is* C^{α} *if* $\boldsymbol{A}(\boldsymbol{x})$ *is. This function is analytical everywhere except for the finite number of points in the complex plane* $a_i = \lambda_i$*.*

Proof.

All eigenvalues of the positive definite matrix $\boldsymbol{A}(\boldsymbol{x})$ are real and positive. Denote by $\rho(\boldsymbol{x})$ the minimal positive eigenvalue of the matrix $\boldsymbol{A}(\boldsymbol{x})$. The function $\rho(\boldsymbol{x})$ is continuous in closure $\bar{\mathbb{G}}$ of $\mathbb{G}$ and not equal to zero. This follows from the fact that the rank of the matrix $\boldsymbol{A}(\boldsymbol{x})$ is constant. Hence, $\rho_0 = \min_{\boldsymbol{x}\in\bar{\mathbb{G}}} \rho(\boldsymbol{x}) > 0$. In the complex plane $\mathbb{Z}$, consider a closed loop Γ that lies in the right half-plane and contains inside itself all positive eigenvalues of the matrices $\boldsymbol{A}(\boldsymbol{x})$. For this, the loop Γ must intersect the real axis to the left of the point ρ_0. Note that matrix $\boldsymbol{A}(\boldsymbol{x}) - z\,\boldsymbol{I})$ does not degenerate on this loop, therefore, we conclude that the elements of the matrix $\boldsymbol{B}(\boldsymbol{x})$ have the same smoothness in the parameter $\boldsymbol{x} \in \bar{\mathbb{G}}$, as do the elements of the matrix $\boldsymbol{A}(\boldsymbol{x})$. To be sure, note that unless the matrix $\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x}) - z\,\boldsymbol{I}$ degenerates in a neighborhood of some point $\boldsymbol{x}_0$, $\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{C}(\boldsymbol{x}_0)[\boldsymbol{C}^{-1}(\boldsymbol{x}_0)\boldsymbol{C}(\boldsymbol{x})]$ and $\boldsymbol{C}^{-1}(\boldsymbol{x}) = [\boldsymbol{C}^{-1}(\boldsymbol{x}_0)\boldsymbol{C}(\boldsymbol{x})]^{-1}\boldsymbol{C}^{-1}(\boldsymbol{x}_0)$. Hence for sufficiently small $\mid \boldsymbol{x} - \boldsymbol{x}_0 \mid$, the matrix $\boldsymbol{C}^{-1}(\boldsymbol{x}_0)\boldsymbol{C}(\boldsymbol{x})$ is close to the identity matrix, and therefore, the inverse matrix may be written as follows

$$[\boldsymbol{C}^{-1}(\boldsymbol{x}_0)\boldsymbol{C}(\boldsymbol{x})]^{-1} = \sum_{k=0}^{\infty}(\boldsymbol{C}^{-1}(\boldsymbol{x}_0)\boldsymbol{C}(\boldsymbol{x}) - I)^k,$$

which converges in the norm for $\mid \boldsymbol{x}-\boldsymbol{x}_0 \mid$. Hence, $\boldsymbol{C}^{-1}(\boldsymbol{x}) = (\boldsymbol{A}(\boldsymbol{x})-z\,\boldsymbol{I})^{-1} = \sum(\boldsymbol{C}^{-1}(\boldsymbol{x}_0)\boldsymbol{C}(\boldsymbol{x})- I)^k\boldsymbol{C}^{-1}(\boldsymbol{x}_0)$. The last formula implies that $\boldsymbol{C}^{-1}(\boldsymbol{x})$ has the same smoothness in the parameter $\boldsymbol{x}$ as $\boldsymbol{A}(\boldsymbol{x})$ does. Consequently $\boldsymbol{B}(\boldsymbol{x})$ is C^{α}, whenever $\boldsymbol{A}(\boldsymbol{x})$ is.

To complete the proof it remains to show that $\boldsymbol{B}(\boldsymbol{x})^m = \boldsymbol{A}(\boldsymbol{x})$. To see that, note that using the eigen decomposition

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{U}\,\boldsymbol{\Lambda}\,\boldsymbol{U}',$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix that contains the eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ on the main diagonal, and $\boldsymbol{U}$ is an orthogonal matrix. Note also that $\boldsymbol{U}$ does not depend on z and could be integrated out in (**??**), therefore

$$\boldsymbol{B}(\boldsymbol{x}) = \frac{1}{2\pi\, i}\int_{\Gamma} f(z)dz = \frac{1}{2\pi\, i}\int_{\Gamma} \sqrt{z}(\boldsymbol{A}(\boldsymbol{x}) - z\,\boldsymbol{I})^{-1}dz = \boldsymbol{U}\left[\frac{1}{2\pi\, i}\int_{\Gamma}{}' \begin{pmatrix} \frac{\sqrt{z}}{\lambda_1 - z} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\sqrt{z}}{\lambda_1 - z} \end{pmatrix} dz\right]\boldsymbol{U}', \tag{9}$$

now applying the Cauchy integral formula,

$$\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{U} \begin{pmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_n} \end{pmatrix} \boldsymbol{U}', \tag{10}$$

and clearly $\boldsymbol{B}(\boldsymbol{x})^2 = \boldsymbol{A}(\boldsymbol{x})$ ■

Now we proceed with the proof of Theorem 1.

Proof. Set $\boldsymbol{A}(\boldsymbol{x}, \epsilon) = \boldsymbol{A}(\boldsymbol{x}) + \epsilon \boldsymbol{I}$ for $\epsilon \in (0, 1)$. Lemma 1 implies that there exists a continuous positive definite matrix-value function $\boldsymbol{B}(\boldsymbol{x}, \epsilon)$ such that $\boldsymbol{B}(\boldsymbol{x}, \epsilon)^m = \boldsymbol{A}(\boldsymbol{x}, \epsilon)$ and $\boldsymbol{B}(\boldsymbol{x}, \epsilon)$ is continuous differentiable with respect to $\epsilon \in (0, 1)$. Thus we have

$$\boldsymbol{I} = \frac{\partial B}{\partial \epsilon}(\boldsymbol{x}, \epsilon)\, \boldsymbol{B}(\boldsymbol{x}, \epsilon)^{m-1} + \boldsymbol{B}(\boldsymbol{x}, \epsilon) \frac{\partial B}{\partial \epsilon} + \cdots + \boldsymbol{B}(\boldsymbol{x}, \epsilon)^{m-1} \frac{\partial B}{\partial \epsilon}(\boldsymbol{x}, \epsilon). \tag{11}$$

For fixed $(\boldsymbol{x}, \epsilon)$, choose an orthogonal matrix $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{x}, \epsilon)$ such that

$$\boldsymbol{U}\, \boldsymbol{B}(\boldsymbol{x}, \epsilon) \boldsymbol{U}' = diag(\lambda_1, \dots, \lambda_n) = \Lambda,$$

and denote $\Theta = \boldsymbol{U} \frac{\partial B}{\partial \epsilon}(\boldsymbol{x}, \epsilon) \boldsymbol{U}'$. Then (11) implies

$$\boldsymbol{I} = \boldsymbol{U}\boldsymbol{I}\boldsymbol{U}' = \Theta \Lambda^{m-1} + \Lambda \Theta \Lambda^{m-2} + \cdots + \Lambda^{m-1} \Theta,$$

hence

$$\boldsymbol{I} = \theta_{ij} (\lambda_j^{m-1} + \lambda_j^{m-2} \lambda_i + \cdots + \lambda_i^{m-1}),$$

which implies

$$\frac{\partial B}{\partial \epsilon}(\boldsymbol{x}, \epsilon) = \boldsymbol{U} diag \left(\frac{1}{m\, \lambda_1^{m-1}}, \dots, \frac{1}{m\, \lambda_n^{m-1}} \right) \boldsymbol{U}',$$

because $\theta_{ij} = 0$ for all $i \neq j$ and $\theta_{ii} = 1/(m\, \lambda_i^{m-1})$. Therefore, for $i, j = 1, \dots, n$

$$| \frac{\partial B}{\partial \epsilon}(\boldsymbol{x}, \epsilon) | \leq \frac{1}{m} \epsilon^{1/(m-1)},$$

where we use $| \boldsymbol{U}' | = 1/(| \boldsymbol{U} |)$, such that

$$| \frac{\partial B}{\partial \epsilon}(\boldsymbol{x}, \epsilon) | \leq | \boldsymbol{U} |\, | \Theta |\, | \boldsymbol{U}' | \leq \frac{1}{m} \epsilon^{1/(m-1)},$$

and the last inequality follows from the fact that $\lambda_i \geq \epsilon^{1/m}$ for $i = 1, \ldots, n$. Therefore, Arzelà theorem applies and we could choose a subsequence $\boldsymbol{B}(\boldsymbol{x}, \epsilon)_1$ that converges to the limit $\boldsymbol{B}(\boldsymbol{x})$ as $\epsilon \to 0$. So, $\{\boldsymbol{B}(\boldsymbol{x}, \epsilon)\}$ uniformly converges on $\mathbb{G}$ as $\epsilon \to 0$ and $\boldsymbol{B}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \boldsymbol{B}(\boldsymbol{x}, \epsilon) \geq 0$ satisfies that $\mid \boldsymbol{B}(\boldsymbol{x}, \epsilon) - \boldsymbol{B}(\boldsymbol{x}) \mid \leq \epsilon^{1/m}$ and $\boldsymbol{B}(\boldsymbol{x})^m = \boldsymbol{A}(\boldsymbol{x})$, which induce that $\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x})^{1/m}$ is continuous on $\mathbb{G}$ ■

Another important result that is shown in Chen and Huan (1997) is the following

Theorem 2 *Let $\boldsymbol{A}(\boldsymbol{x})$ be a $C^{\alpha}(\mathbb{E})$ non-negative definite matrix-valued function for an open domain $\mathbb{E} \subset \mathbb{R}^N$ for some $\alpha \in (0, 2]$. then the m-th root function $\boldsymbol{A}^{1/m}(\boldsymbol{x})$ is in $C^{\alpha/m}(\mathbb{E})$.*

Our goal is to adopt these results for econometric applications that use matrix-valued functions of the covariance matrix. The following result is a trivial adaptation of previous results for this case.

Corollary 1 *Let $\boldsymbol{A}(\boldsymbol{X})$ be a matrix-valued functional that depends on a symmetric non-negative definite matrix $\boldsymbol{X}$. Define $\boldsymbol{x} = vech(\boldsymbol{X})$ such that $\boldsymbol{x} \in \mathbb{G} \subset \mathbb{R}^N$, then $\boldsymbol{A}(\boldsymbol{X})$ can be written as a functional of the vector, $\boldsymbol{x}$, $\boldsymbol{A}(\boldsymbol{X}) = \widetilde{\boldsymbol{A}(\boldsymbol{x})}$. Let $\boldsymbol{A}(\boldsymbol{X})$ be continuous on $\mathbb{G}$ if $\widetilde{\boldsymbol{A}(\boldsymbol{x})}$ is continuous on $\mathbb{G}$. Therefore, if $\boldsymbol{A}(\boldsymbol{X})$ is continuous on $\mathbb{G}$. then so is the m-th root functional $\boldsymbol{A}^{1/m}(\boldsymbol{X})$ of $\boldsymbol{A}(\boldsymbol{X})$. Thus, if $\boldsymbol{A}(\boldsymbol{X})$ is a $C^2(\mathbb{E})$ non-negative definite matrix-valued function for an open domain $\mathbb{E} \subset \mathbb{R}^N$, so $\boldsymbol{A}^{1/m}(\boldsymbol{X})$ is in $C^1(\mathbb{E})$.*

Remark 5 *For $\boldsymbol{A}(\boldsymbol{X}) = \boldsymbol{X}$, symmetric positive semi-definite matrix $\boldsymbol{X}$ and $m = 2$, the last theorem states that the square root of a symmetric positive semi-definite matrix is a continuous transformation in the elements of the matrix and has continuous derivatives.*

Therefore, proposition 1 immediately follows. Proposition 2 follows because convergence in probability is preserved for continuous functions.

7.1 Uniform convergence of the bootstrap approximations CDF of W_I

We start with the univariate normal distribution to explain the intuition and then generalize the result to the multivariate normal distribution. Consider the *probability density function* (*pdf*) of a univariate normal distribution $N(a, \sigma^2)$

$$f_{\xi}(x, a, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, exp - \frac{(x-a)^2}{2\sigma^2}, \qquad x \in \mathbb{R}. \tag{12}$$

Now denote consistent estimates of the parameters a and σ by a_T and σ_T. Hence, $\sigma_T \underset{p}{\longrightarrow} \sigma$ and $a_T \underset{p}{\longrightarrow} a$. Note that $f_{\xi}(x, a, \sigma)$ is continuous at all $x \in \mathrm{R}$ for every $\sigma > 0$ and for any value of a. Define $f_{\xi_T}(x, a_T, \sigma_T)$ as the *pdf* of ξ with parameters a and σ substituted by the respective estimates. Due to continuity of $f_{\xi}(x, a, \sigma)$, the following pointwise convergence in probability of the sequence of *pdf*$'s$, $f_{\xi_T}(x, a_T, \sigma_T)$ holds

$$f_{\xi_T}(x, a_T, \sigma_T) \underset{p}{\longrightarrow} f_{\xi}(x, a, \sigma), \qquad \forall x \in \mathbb{R}. \tag{13}$$

Denote the *cumulative density function* (*cdf*) of this distribution

$$F_{\xi}(x, a, \sigma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma}\, exp\frac{(v-a)^2}{2\sigma^2} dv,$$

and $F_{\xi_T}(x, a_T, \sigma_T)$ similarly defined.

Then the same pointwise convergence result holds for a sequence of the *cdf*$'s$ because every element of the sequence is continuous in a_T and σ_T

$$F_{\xi_T}(x, a_T, \sigma_T) \underset{p}{\longrightarrow} F_{\xi}(x, a, \sigma), \qquad \forall x \in \mathbb{R}. \tag{14}$$

Denote $F_{\xi_T}(x, a_T, \sigma_T) = F_T(x)$ for ease of exposition. Then, due to boundedness, continuity and monotonicity of *cdf* pointwise convergence implies uniform convergence for this sequence from arguments similar to the proof of ***Glivenko-Cantelli theorem*** (e.g. Davidson, 1994, 21.5).

Lemma 2 *If* $F_T(x) \underset{p}{\longrightarrow} F(x)$ *pointwise, for* $x \in R$, *then* $\underset{x}{sup}|F_T(x) - F(x)| \underset{p}{\longrightarrow} 0$.

Denote by $G_T(x)$ and $G(x)$ the *cdf*$'s$ of ξ^2, note that $G_T(x)$ is a function of the estimated values of a and σ. Then similar result holds for $G_T(x)$ and $G_(x)$.

Lemma 3 $\underset{x}{sup}|G_T(x) - G(x)| \underset{p}{\longrightarrow} 0$.

It follows from the continuous mapping theorem for ξ. Therefore, the parametric bootstrap is asymptotically consistent if the approximated distribution is univariate normal.

First, the extension to the multivariate normal is trivial if the variance covariance matrix is of full rank. The domain of $F_T(\boldsymbol{x})$ to $F(\boldsymbol{x})$ is in $\mathbb{R}^k$, where k is the dimension of $\boldsymbol{x}$, and the distribution of $\boldsymbol{\xi}$ depends on the vector of means $\boldsymbol{a}$ and covariance matrix $\boldsymbol{\Sigma}$. Despite that there is no close form expression for the *cdf*, it could be described as a multiple integral of the *pdf* These integrals are continuous functions with respect to the estimates of the covariance matrix and the mean vector, so the continuity argument establishing pointwise convergence in probability of $F_T(\boldsymbol{x})$ to $F(\boldsymbol{x})$ holds. Thus, by similar arguments as above, uniform convergence follows for *cdf*$'s$ of $W_I = \boldsymbol{\xi}'\boldsymbol{\xi}$, $G_T(x)$ to $G(x)$.

Consider now a vector of multivariate normal random variables, $\boldsymbol{\xi}_T$, with zero vector of means and a rank deficient variance-covariance matrix, $\boldsymbol{\Sigma}_T$. There is a square root of this matrix, $\boldsymbol{\Sigma}_T^{1/2}$. For a given vector of random variables $\boldsymbol{\xi}_T$, there is a transformation of another vector of multivariate normal random variables, $\boldsymbol{\phi}_T$, with an identity covariance matrix and mean vector of zeros, such that is $\boldsymbol{\xi}_T = \boldsymbol{\Sigma}^{1/2}\,\boldsymbol{\phi}_T$.

Denote the probability measure of $\boldsymbol{\phi}$ by $\lambda_{\boldsymbol{\phi}}(\boldsymbol{u}, \boldsymbol{I})$. Then,

$$P(\boldsymbol{\xi} < \boldsymbol{x}_0) = P(\boldsymbol{\Sigma}^{1/2}\,\boldsymbol{\phi} < \boldsymbol{x}_0) = \int \mathbb{1}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{u} < \boldsymbol{x}_0)\, \mathrm{d}\lambda_{\boldsymbol{\phi}}. \tag{15}$$

This function is continuous with respect to $\boldsymbol{\Sigma}^{1/2}$ because the sequence of indicator functions $\mathbb{1}(\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{u} < \boldsymbol{x}_0) \to \mathbb{1}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{u} < \boldsymbol{x}_0)[\lambda_{\boldsymbol{\phi}}]a.e.$ if $\boldsymbol{\Sigma}_k^{1/2} \to \boldsymbol{\Sigma}^{1/2}$, and dominated convergence theorem (e.g. Davidson, 1994, 4.12) applies. Thus,

$$P(\boldsymbol{\Sigma}_k^{1/2}\,\boldsymbol{y} < \boldsymbol{x}_0) \to P(\boldsymbol{\Sigma}^{1/2}\,\boldsymbol{y} < \boldsymbol{x}_0) \tag{16}$$

Therefore, is a $P(\boldsymbol{\Sigma}^{1/2}\,\boldsymbol{y} < \boldsymbol{x}_0)$ continuous function with respect to $\boldsymbol{\Sigma}^{1/2}$ and this case is similar to previous cases. In this case the *cdf* of W_I is $P(\boldsymbol{\xi}'\boldsymbol{\xi} < x) = P(\boldsymbol{\phi}\,\boldsymbol{\Sigma}\,\boldsymbol{\phi} < x)$. Similar arguments as in case of $P(\boldsymbol{\xi} < \boldsymbol{x}_0)$ apply and deliver the uniform convergence in probability.

Note that one can extend the above results to any differentiable transformation G with uniformly bounded derivative and obtain the uniform convergence of $G(F(\boldsymbol{x}, \boldsymbol{a}_T, \boldsymbol{\Sigma}_T^2))$ to $G(F(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{\Sigma}^2))$. This is a specific case of a more general result establishing consistency of the bootstrap (e.g. Horowitz, 2001, Theorem 2.1). For the case at hand it is also sufficient to rely on the convergence of distribution of $\boldsymbol{\xi}_T$ to $\boldsymbol{\xi}$ and consequently on the continuous mapping theorem to get the pointwise convergence of the sequence of $cdf's$ of $\boldsymbol{\xi}_T'\boldsymbol{\xi}_T$ to the *cdf* of $\boldsymbol{\xi}'\boldsymbol{\xi}$.

To prove the consistency of the residual bootstrap consider the empirical distribution function(edf) to approximate the *cdf* of the normal distribution. Consistency of the estimators of the structural parameters together with a law of large numbers delivers pointwise convergence of the sequence of $edf's$ to the *cdf*. Glivenko-Cantelli theorem implies uniform convergence. Then, similar to the parametric bootstrap case, the continuous mapping theorem ensures that the asymptotic distribution of W_I is approximated consistently by the residual bootstrap.

7.2 Illustration of continuity of a square-root of a matrix

Let us consider the example from Dufour et al. (2011) to illustrate how the square root of the matrix is calculated in that case and its continuity shown.

Example 1 *Let $\boldsymbol{A}(x)$ be the matrix function defined as:*

$$\boldsymbol{A}(x) = \begin{cases} \begin{pmatrix} 1+x & 0 \\ 0 & 1-x \end{pmatrix}, & \text{if } x < 0; \\ \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix}, & \text{if } x \geqslant 0. \end{cases} \tag{17}$$

This matrix function is continuous at $x = 0$, with $\boldsymbol{A}(0) = \boldsymbol{I}_2$, however the eigenvectors differ for $x \to 0^+$ and $x \to 0^-$. If $x \to 0^+$, the eigenvectors are $\frac{1}{\sqrt{2}}(1;1)'$ and $\frac{1}{\sqrt{2}}(1;-1)'$, whereas for $x \to 0^-$ the eigenvectors are $(1;0)'$ and $(0;1)'$. Note that the eigenvectors are orthonormal, and yet they are not continuous with respect to the elements of the matrix $\boldsymbol{A}(x)$.

Now consider the square-root of the matrix $\boldsymbol{A}(x)$ defined according to Freidlin (1968)

$$\boldsymbol{B}(x) = \frac{1}{2\pi\,i}\int_\Gamma f(z)dz = \frac{1}{2\pi\,i}\int_\Gamma \sqrt{z}(\boldsymbol{A}(x) - z\,\boldsymbol{I}_2)^{-1}dz, \tag{18}$$

where Γ is a closed contour that contains all non-zero eigenvalues of $\boldsymbol{A}(x)$, and is in the right complex half plane $\{z : \text{Re } z > 0\}$. This function is analytical everywhere except for the finite number of points in the complex plane $a_i = \lambda_i$. In this case $\lambda_1 = (1+x)$ and $\lambda_2 = (1-x)$, thus using the Cauchy's residue theorem (see, e.g., Knopp, 1996)

$$\boldsymbol{B}(x) = 2\pi\, i \sum_{i=1}^{dim(\boldsymbol{A}(x))} Res(f(a_i)), \tag{19}$$

where $Res(f(a_i))$ is the residue of $f(z)$ at a pole a_i, and is calculated as follows

$$Res(f(a_i)) = \lim_{z \to a_i} (z - a_i) \times \sqrt{z} \times (\boldsymbol{A}(x) - z\, \boldsymbol{I}_2)^{-1}. \tag{20}$$

Then, for $x < 0$

$$Res(f(1+x)) = \lim_{z \to 1+x} (z - 1 - x) \times \sqrt{1+x} \times \frac{1}{(1-z+x)(1-z-x)} \begin{pmatrix} 1-x-z & 0 \\ 0 & 1+x-z \end{pmatrix} = \begin{pmatrix} \sqrt{1+x} & 0 \\ 0 & 0 \end{pmatrix}.$$

Similarly,

$$Res(f(1-x)) = \lim_{z \to 1-x} (z - 1 + x) \times \sqrt{1-x} \times \frac{1}{(1-z+x)(1-z-x)} \begin{pmatrix} 1-x-z & 0 \\ 0 & 1+x-z \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{1-x} \end{pmatrix},$$

whereas for $x \geqslant 0$

$$Res(f(1+x)) = \lim_{z \to 1+x} (z - 1 - x) \times \sqrt{1+x} \times \frac{1}{(1-z+x)(1-z-x)} \begin{pmatrix} 1-z & -x \\ -x & 1-z \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{1+x}}{2} & \frac{\sqrt{1+x}}{2} \\ \frac{\sqrt{1+x}}{2} & \frac{\sqrt{1+x}}{2} \end{pmatrix}.$$

Likewise,

$$Res(f(1-x)) = \lim_{z \to 1-x} (z - 1 + x) \times \sqrt{1-x} \times \frac{1}{(1-z+x)(1-z-x)} \begin{pmatrix} 1-z & -x \\ -x & 1-z \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{1-x}}{2} & -\frac{\sqrt{1-x}}{2} \\ -\frac{\sqrt{1-x}}{2} & \frac{\sqrt{1-x}}{2} \end{pmatrix}.$$

Therefore,

$$\boldsymbol{B}(x) = \begin{cases} \begin{pmatrix} \sqrt{1+x} & 0 \\ 0 & \sqrt{1-x} \end{pmatrix} & \text{if } x < 0; \\ \begin{pmatrix} \frac{\sqrt{1+x}}{2} & \frac{\sqrt{1+x}}{2} \\ \frac{\sqrt{1+x}}{2} & \frac{\sqrt{1+x}}{2} \end{pmatrix} + \begin{pmatrix} \frac{\sqrt{1-x}}{2} & -\frac{\sqrt{1-x}}{2} \\ -\frac{\sqrt{1-x}}{2} & \frac{\sqrt{1-x}}{2} \end{pmatrix} & \text{if } x \geqslant 0. \end{cases} \tag{21}$$

Thus, $\lim_{x \to 0^+} B(x) = \lim_{x \to 0^-} \boldsymbol{B}(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.